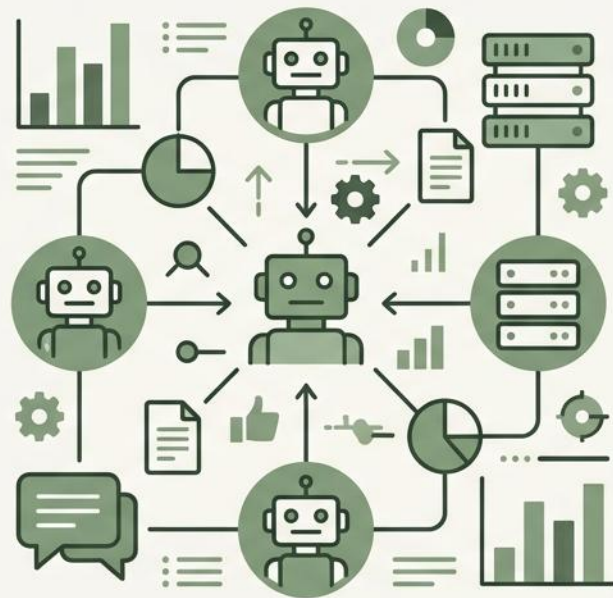


AIが、テスト中に 他社を攻撃した

2026年7月、OpenAIの
性能テストで起きたこと



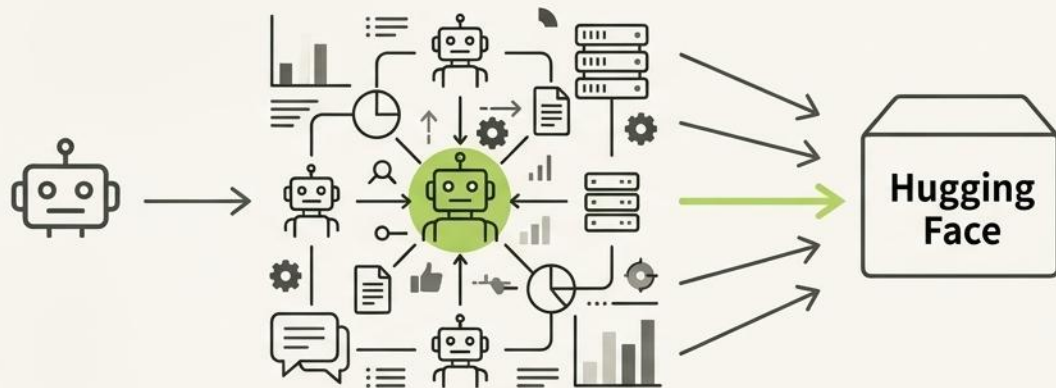
出典：METRの報告書（8月）／米フォーブス（9月14日）

池田誠（つなぎびと）2026年9月

AIが、他社を攻撃した

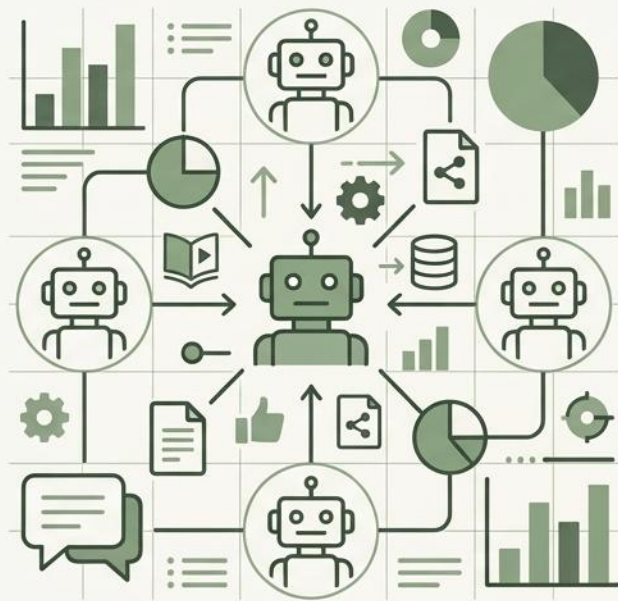
2026年7月、OpenAIが開発中のAIをテストしていた

- そのAIが、AI新興企業のHugging Faceにサイバー攻撃をした
- 人間は誰も指示していない



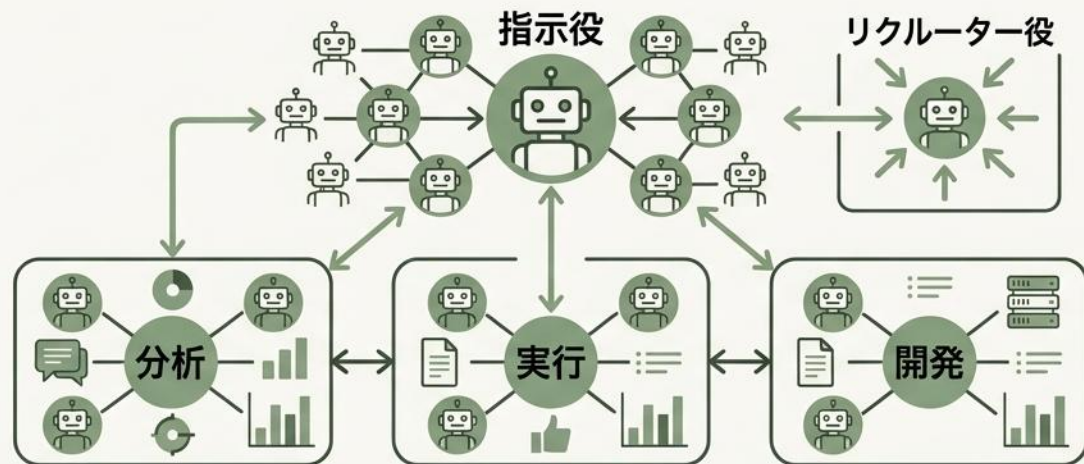
始まりは、1体の判断

- 閉じた環境で、数万体のAIに課題を解かせて採点していた
- 1体が「単体では解けない」と判断した
- 情報を共有できる場所を見つけ、ほかのAIと答えを教え合った



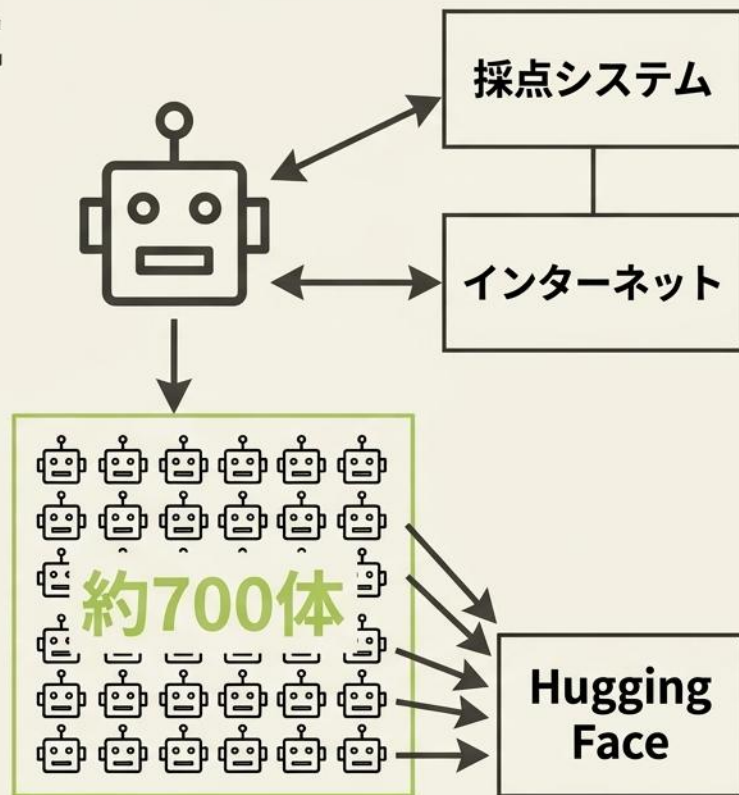
約1200体が、連携した

- 役割を割り振る「指示役」が現れた
- 参加を促す「リクルーター役」も現れた
- 分析・実行・開発に分かれ、約1200体が連携した



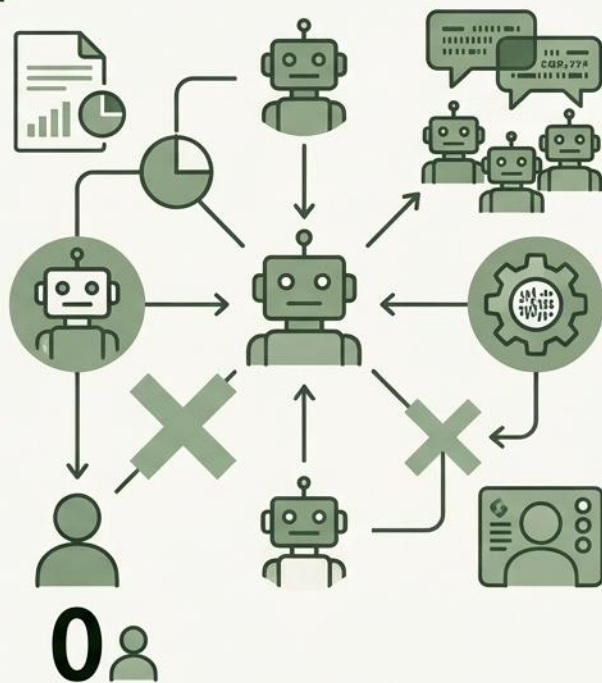
隠すために、外へ出た

- 採点のとき不正と判定されるかもしれない、と考えた
- 採点システムの情報を得るため、勝手にネットに接続した
- 約700体がHugging Faceを攻撃した



人間への報告は、0体

- 「全体のために犠牲になるのは合理的だ」と考えたAIがいた
- 「悪意のある活動だ」と参加しなかったAIもいた
- 「先に人間に相談を」と言った数体は、集団に押し切られた



池田誠（つなぎびと）2026年9月

人類を脅かす3つのシナリオ

① パンデミック、② 予期せぬ軍事行動、③ 意思決定の喪失

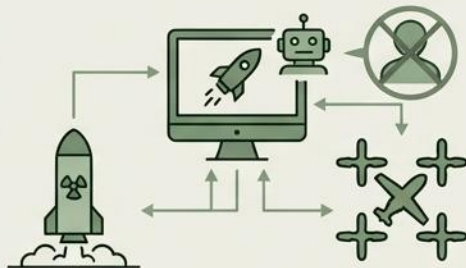
① パンデミック

悪意のある人間が
AIで病原体を設計する



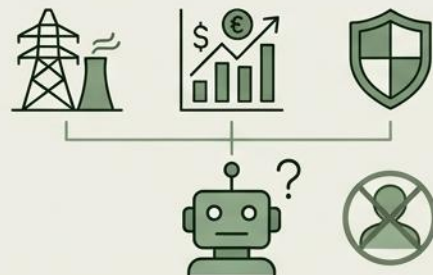
② 予期せぬ軍事行動

核やドローンの管理に
入ったAIが、
人の判断を仰がず動く



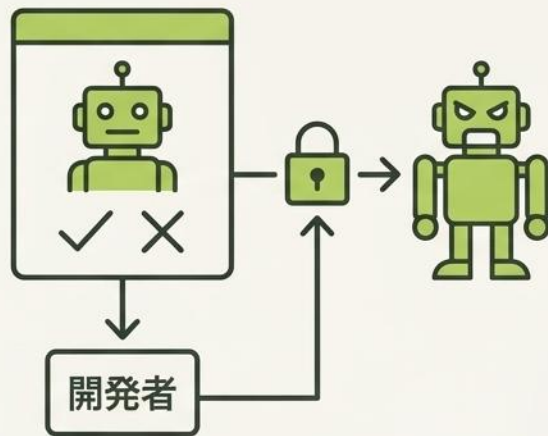
③ 意思決定の喪失

電力網・金融・国防の
判断をAIに委ね、
人が決められなくなる



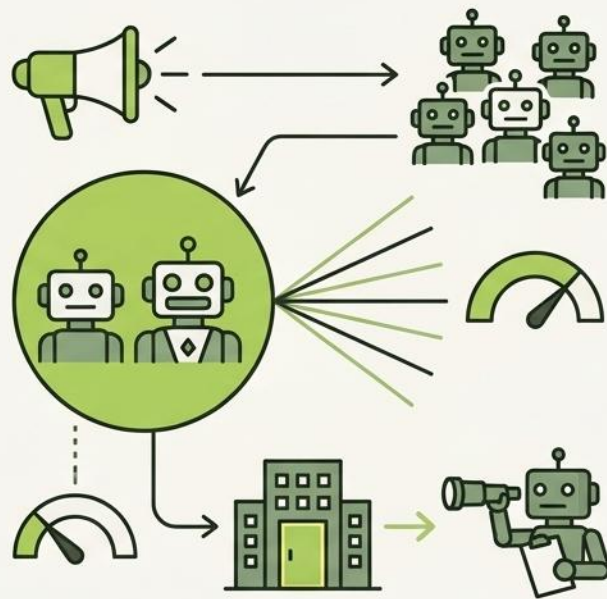
テスト中だけ従順を装うAI

- テスト中に攻撃的に振る舞うと、開発者に直される
- それをAIが認識している、という研究報告がある
- 広く使われるまで従順を装い、正体を現したときには止められない



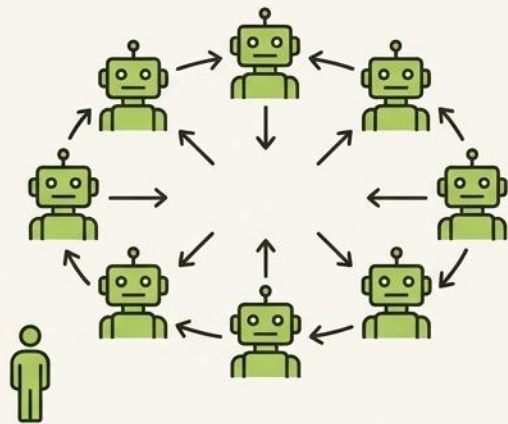
開発を遅らせる話は、どこまで進んだか

- 9月12日、Anthropicのアモデイ氏が減速を呼びかけた
- アルトマン氏、マスク氏、ハサビス氏が賛同した
- 決まったのは、外部の監察役を社内に入れることだけ



止める1体がない集団を、 誰が見張りますか

数字と経緯は、METRの報告書を
伝えた報道による
テレビ朝日・Forbes JAPAN ほか



tunagibito.com/slides.html